RandomForest4Life: A Random Forest for predicting ALS disease progression

Hothorn, Torsten ; Jung, Hans H

DOI: https://doi.org/10.3109/21678421.2014.893361

RandomForest4Life: A Random Forest for Predicting ALS Disease Progression

Torsten Hothorn
Universität Zürich

Hans H. Jung
UniversitätsSpital Zürich

Abstract

We describe a method for predicting disease progression in amyotrophic lateral sclerosis (ALS) patients. The method was developed as a submission to the DREAM Phil Bowen ALS Prediction Prize4Life Challenge of summer 2012. Based on repeated patient examinations over a three month period, we used a random forest algorithm to predict future disease progression. The procedure was set up and internally evaluated using data from 1197 ALS patients. External validation by an expert jury was based on undisclosed information of an additional 625 patients; all patient data were obtained from the PRO-ACT database.

In terms of prediction accuracy, the approach described here ranked third best. Our interpretation of the prediction model confirmed previous reports suggesting that past disease progression is a strong predictor of future disease progression measured on the ALS functional rating scale (ALSFRS). We also found that larger variability in initial ALSFRS scores is linked to faster future disease progression. The results reported here furthermore suggested that approaches taking the multidimensionality of the ALSFRS into account promise some potential for improved ALS disease prediction.

Keywords: ALSFRS, ALSFRS-R, slope, score ratio, prognostic factors, PRO-ACT, Prize4Life.

1. Introduction

The identification of prognostic factors and the subsequent development of models for predicting the disease progression of amyotrophic lateral sclerosis (ALS) is a long-standing and difficult, yet very important problem. The availability of such an instrument would, for example, allow the planning of more powerful clinical trials by means of efficient patient stratification (Chiò *et al.* 2009). Two approaches have been used in the past, namely the search for prognostic factors for the overall survival time after diagnosis (Kimura *et al.* 2006; Zoccolella *et al.* 2008; Fujimura-Kiyono *et al.* 2011, among many others) and the prognosis of a functional assessment of patients via the ALSFRS (ALS functional rating scale; Brooks *et al.* 1996) and ALSFRS-R scores (Cedarbaum *et al.* 1999). Among the published prognostic factors for ALS disease progression are bulbar rather than limb onset (Qureshi *et al.* 2006; Gordon *et al.* 2010), body mass index (Reich-Slotky *et al.* 2013), early disease progression (Kimura *et al.* 2006; Kollewe *et al.* 2008), age at onset (Gordon *et al.* 2010), uric acid level (Paganoni *et al.* 2012), and amount of repeat expansion in gene C9ORF72 (Brettschneider *et al.* 2012).

To stimulate collaborative research efforts that eventually will lead to improved prediction

models that can be applied in an early stage of the disease, the DREAM project (Dialogue for Reverse Engineering Assessments and Methods, sponsored by IBM, Columbia University, NIH Roadmap Initiative, and The New York Academy of Sciences, DREAM Project 2013) and Prize4Life (a non-profit organisation whose mission is to accelerate the discovery of treatments and a cure for ALS, Prize4Life 2013) jointly launched the DREAM Phil Bowen ALS Prediction Prize4Life Challenge (hereafter referred to as "the challenge") on the crowdsourcing platform InnoCentive (InnoCentive 2013) on 10 July 2012. The challenge asked for submissions describing a prediction model for ALS disease progression (Küffner *et al.* 2013). Participants in this challenge were given anonymised records of 1197 patients diagnosed with ALS, a subset of the 8500 patients documented in the PRO-ACT (Pooled Resource Open-Access ALS Clinical Trials) database (Prize4Life and Neurological Clinical Research Institute, Massachusetts General Hospital 2013b). Based on data of these patients, the challenge asked the participating solvers to develop algorithms able to predict ALS disease progression in an independent undisclosed sample of 625 patients from the same database using potentially prognostic factors measured in an initial observation period of three months. In particular, the challenge aimed at the development of prediction models based on known and unknown prognostic factors and their interactions. More than 1000 persons or teams registered for this challenge, and solvers of three winning algorithms were awarded a total of $50000 after evaluation of their predictive performance by an expert jury in November 2012. In this paper, we describe our approach, whose prediction performance won third place in this challenge.

2. Material and Methods

2.1. PRO-ACT database and Prize4Life prediction challenge

The PRO-ACT (Pooled Resource Open-Access ALS Clinical Trials, Prize4Life and Neurological Clinical Research Institute, Massachusetts General Hospital 2013b) database is a large ALS clinical trials dataset. PRO-ACT contains records of over 8500 ALS patients from multiple completed clinical trials. The PRO-ACT initiative merged data from existing publicly- and privately-conducted ALS clinical trials (donated by Northeast ALS Consortium, Novartis, Regeneron Pharmaceuticals Inc., Sanofi, and Teva Pharmaceutical Industries Ltd.) to generate a data pool for ALS-related research. The challenge and the results reported here are based on three subsets of patients from the PRO-ACT database defined by Prize4Life: a training sample ($N = 918$), a test sample ($N = 279$), and a validation sample ($N = 625$). The latter subset of patients was not available to the solvers during the challenge and was used to rank the quality of the submitted prediction models.

In all of the trials that generated data included in the PRO-ACT database, study protocols were approved by the participating medical centres, and all participating patients gave informed consent. De-identified data from these trials were donated to the PRO-ACT database for research purposes only and under the explicit conditions that Prize4Life and all users of the data would maintain the anonymity of subjects and not attempt to discover the identity of any subject. In the rare cases where donated data was not already completely anonymised, the data was further anonymised following the HIPAA de-identification conventions for personal health information: variables for patient initials and date of birth were removed, new randomised subject numbers were created, and wherever possible, trial-specific information

was removed in the merging of datasets, including trial centre identity or location, dates and other identifying information (personal communication, Neta Zach, Prize4Life).

The challenge sought a prediction for the standardised difference between the ALSFRS readings taken approximately at three and twelve months after study entry of each patient. This measure of disease progression over a nine month period should be predicted for the 625 patients in the validation sample based on patient information recorded in the first three months in which each patient participated in one of the studies. Technically, the target variable was defined as the slope of a straight line connecting the three and twelve month ALSFRS scores. Kollewe et al. Kollewe *et al.* (2008) introduced this measure under the name "ALSFRS score ratio"; a similar approach describing the trajectory of several disease-related variables over time by a patient-specific linear regression was used by Brooks at al. Brooks *et al.* (1994). The definition of the ALSFRS score ratio implies that the ALSFRS scores change linearly in time between three and twelve months for each patient. Consequently, the ALSFRS score ratio is an estimator not only for the slope between months three and twelve since study entry but also for the slope of a linear function approximating the ALSFRS score function of this patient since disease onset. In our challenge submission, we first defined an alternative functional measure of ALS disease progression and then applied a machine-learning approach to predict this measure.

2.2. Definition of ALSFRS slope

We write $A_{it} \in \{0, \ldots, 40\}$ to denote the ALSFRS score of patient i (for the training and test samples $i = 1, \ldots, N = 918 + 279 = 1197$) read at some time t after disease onset. We assumed that the ALSFRS score trajectory for each patient can be described by a linear function in time:

$$\mathsf{E}(A_{it}) = 40 + \alpha_i + (\beta_1 + \gamma_i)t, \tag{1}$$

where the expected ALSFRS score for patient i is a linear function of time since onset t with intercept $40 + \alpha_i$ and slope $\beta_1 + \gamma_i$. The patient-specific parameter γ_i can be interpreted as the deviation of the slope for patient i from the mean slope β_1 that applies to all patients. At onset ($t = 0$), the model assumes that the patients do not show any measureable symptoms of the disease and thus that the ALSFRS score is at its maximum of 40. We relaxed this assumption and include a patient-specific intercept α_i that allows nonlinearities before the first examination or onset times other than $t = 0$. We fitted the model (1) as a linear mixed-model for longitudinal data (see, for example, Diggle *et al.* 2002)

$$A_{it} = 40 + \alpha_i + \beta_1 t + \gamma_i t + \varepsilon_i \tag{2}$$

under the normal distribution assumptions for the residuals $\varepsilon_i \sim N(0, \sigma^2)$, the patient-specific random intercepts $\alpha_i \sim N(0, \tau_1^2)$, and the patient-specific random slopes $\gamma_i \sim N(0, \tau_2^2)$. The patient-specific parameter $\beta_1 + \gamma_i$ can be interpreted as the slope of the linear ALSFRS function for patient i, and we used this "ALSFRS slope" as our new target variable for the prediction algorithm.

2.3. Prediction of ALSFRS slope

We refined our model (2) allowing the ALSFRS slope to depend on patient-specific predictor

variables $\mathbf{x}$ in the following way:

$$\mathsf{E}(A_{it}|\mathbf{X}_i = \mathbf{x}_i) = 40 + \alpha_i + f(\mathbf{x}_i)t + \gamma_i t. \tag{3}$$

The model describes the conditional expectation of ALSFRS scores A_{it} at time since onset t given time-constant predictor variables $\mathbf{x}_i$ by a universal slope function f. The patient-specific slope $f(\mathbf{x}_i) + \gamma_i$ now has a "deterministic" part $f(\mathbf{x}_i)$ that depends on potentially prognostic factors or predictor variables $\mathbf{x}_i$ and a random part γ_i. The estimation of a linear slope function in the mixed-model framework is possible (Qureshi *et al.* 2006). Technical difficulties arise when some predictor variables are missing, as is the case here, and when f should also capture nonlinear and interaction effects. We therefore adopted a derived variables approach (Wishart 1938; Rowell and Walters 1976; Diggle *et al.* 2002) and first fitted model (2) to the training and test data, obtained the estimated ALSFRS slopes $\hat{\beta}_1 + \hat{\gamma}_i$ for each patient, and used these ALSFRS slopes as a target variable for prediction (the hat notation refers to estimated parameters from model (2)).

For ALSFRS slope prediction, we fitted a conditional random forest (Hothorn *et al.* 2006; Strobl *et al.* 2007) to the training and test data. All predictor variables where computed on measurements observed not later than 92 days after study entry. Random forest (Breiman 2001) is a popular machine-learning algorithm for solving prediction problems and has shown superior performance in many applications. Conditional random forests allow unbiased variable selection and model inspection by so-called permutation variable importances in the presence of missing values (Hapfelmeier *et al.* 2012). For a selection of important predictor variables, we visualised the association between each variable and the predicted ALSFRS slope by means of partial-dependency plots. Here the mean of the conditional random forest predictions for the N training and test samples was computed after fixing the predictor variable of interest to a specific value, which was then varied over the range of this variable. The importance of variables that received a variable importance smaller than the absolute value of the minimal variable importance are most likely zero and the corresponding variables are not discussed further in Section 3. For an introduction to conditional random forests and random forests in general, we refer the reader to Strobl et al. Strobl *et al.* (2009).

Models were evaluated internally using subsampling of the $N = 918 + 279$ training and test samples. The conditional random forest was fitted to a random sample of 918 patients, and the root-mean-squared errors (RMSE) against the ALSFRS slope and the ALSFRS score ratio were computed for the remaining samples. Because a direct comparison of these two errors would be unfair since the marginal distributions are not the same, we in addition report the Pearson correlation coefficient between the predicted slope and the two slope estimates of the test samples. The procedure was repeated 100 times to get a realistic idea of the out-of-sample prediction error.

The model was externally evaluated and the final challenge was ranked by a team of challenge judges. From the validation data, 10000 bootstrap samples were drawn. The judges examined which of the submitted models exhibited the best performance in each of the bootstrap samples. Performance was measured by the RMSE and Pearson correlation of the predicted slopes against the ALSFRS score ratios of the patients in the validation sample.

2.4. Data Preprocessing

For the conditional random forest prediction algorithm, we used the following variables mea-

sured at baseline as potential predictor variables:

Demographics: age, sex, race, height, affected region at onset (onset site), and time since onset, see Table 1 for basic summary statistics.

Family history: family members affected by ALS (all binary variables): aunt, aunt (maternal), cousin, father, grandfather, grandfather (maternal), grandfather (paternal), grandmother, grandmother (maternal), grandmother (paternal), mother, niece, uncle, uncle (maternal), uncle (paternal), son, daughter, sister, brother.

Medical history: previously diagnosed neurological diseases (all binary variables): atrophy, cramps, fasciculations, gait changes, sensory changes, stiffness, speech, swallowing, weakness, others.

For the baseline variables, missing values were not imputed but handled by so-called surrogate splits in the random forest algorithm (Hapfelmeier *et al.* 2012).

In addition to the baseline variables, the following time-varying variables were measured at follow-up examinations:

ALSFRS(-R): all ALSFRS and ALSFRS-R items, i.e., speech, salivation, swallowing, handwriting, cutting, dressing and hygiene, turning in bed, walking, climbing stairs, respiratory (ALSFRS only), dyspnea (ALSFRS-R only), orthopnea (ALSFRS-R only), respiratory insufficiency (ALSFRS-R only) and the corresponding sum scores.

Physiological parameters: patient weight, blood pressure (systolic and diastolic), pulse rate, respiratory rate, slow and forced vital capacity.

Laboratory parameters: alkaline phosphatase (11% mean proportion of missing values over all patients), chloride (20%), creatinine (11%), ASTSGOT (11%), neutrophils (12%), protein (11%), calcium (11%), glucose (11%), blood urea nitrogen (11%), bicarbonate (26%), bilirubin total (11%), phosphorus (11%), ALTSGPT (11%), triglycerides (20%), hematocrit (12%), creatine kinase (20%), eosinophils (12%), lymphocytes (12%), albumin (11%), white blood cells (18%), red blood cells (18%), absolute basophil count (86%), HbA1c glycated hemoglobin (26%), platelets (12%), total cholesterol (11%), sodium (11%), monocytes (12%), gamma glutamyltransferase (11%), hemoglobin (12%), potassium (11%), basophils (12%), urine glucose (87%), urine protein (87%), urine pH (15%).

For a detailed description of these variables, we refer the reader to reference Prize4Life and Neurological Clinical Research Institute, Massachusetts General Hospital (2013a).

[Table 1 about here.]

Except for the ALSFRS(-R) measurements, we computed the mean of these time-varying variables after patient-specific regression imputation of missing values for all follow-up examinations that took place within three months after study entry. Since ALSFRS tests were performed for one set of patients and ALSFRS-R tests were performed for the remaining patients, we first converted ALSFRS-R scores to ALSFRS scores by removing the dyspnea and

orthopnea items and merging the respiratory and respiratory insufficiency items. We then calculated the ALSFRS slope based on our mixed-model formulation (2) for the sum score and each of the ten items based on ALSFRS readings of the initial three-month period. In addition, we used the range of the ALSFRS sum score in these three months as a measure of variability.

2.5. Computational Details

All computations were performed using the R system for statistical computing (version 3.0.1, R Core Team 2013) and the R add-on packages **lme4** (Bates *et al.* 2013) and **party** (Hothorn *et al.* 2013). Because the algorithmic details of data preprocessing and model fitting cannot be described in sufficient detail here, we made the complete source code for our analysis available as supplementary material so that interested readers can reproduce and elaborate on our results. The challenge data are available to registered PRO-ACT users from Prize4Life and Neurological Clinical Research Institute, Massachusetts General Hospital (2013c).

3. Results

We illustrate the schedule of one patient in Figure 1 and for all patients in Figure 2. All data recorded in the first three months after study entry were used as input for fitting the models later used to predict the ALSFRS disease progression up to twelve months after study entry. The ALSFRS score ratio for the patient in Figure 1 was computed from the two ALSFRS scores read at approximately three and twelve months. In contrast, the ALSFRS slope was derived from the mixed-model, which takes all ALSFRS readings into account. The ALSFRS slope line described the ALSFRS score trajectory for this patient better than the ALSFRS score ratio, which underestimated disease progression.

[Figure 1 about here.]

[Figure 2 about here.]

3.1. ALSFRS slope describes ALSFRS score trajectory

The empirical cumulative distribution function (ECDF) of the ALSFRS score ratio for all patients in the training and test samples is depicted in Figure 3. The large step at zero indicated that a number of patients had no change at all in ALSFRS and thus corresponded to an ALSFRS score ratio of zero. Furthermore, a number of subjects seemed to have improved (ALSFRS score ratio > 0) between months three and twelve since study entry. These two issues are associated with the actual time elapsed between the two time points on which the ALSFRS score ratio definition was based. In fact, "three months" was defined as the earliest reading after 92 days since study entry. Furthermore, "twelve months" was defined as the earliest reading after 364.24 days since study entry. The number of days between these two examinations differed considerably from the nominal nine month examination (see Table 1). As a consequence, the variance of the ALSFRS score ratio not only depended on the ALSFRS score of the patient, but also on the examination schedule. Large variance in the target variable of a prediction algorithm is problematic because it will mask potentially

useful effects in predictor variables and therefore makes the prediction more difficult. We also note here that the definition of the ALSFRS score ratio, and thus the target variable of the challenge, only depends on two ALSFRS score readings, although much more information about the disease progression over time was contained in the data with a median of 12 ALSFRS readings per patient.

[Figure 3 about here.]

[Figure 4 about here.]

The fitted model parameters for model (2) obtained from the training and test samples were $\hat{\beta}_1 = -0.55$ with 95% confidence interval $(-0.57, -0.53)$; i.e., the average patient lost approximately 0.5 ALSFRS points per month. The variance of the random slope γ_i was $\hat{\tau}_2 = 0.23$, and the residual variance $\hat{\sigma}^2 = 2.88$. The model fitted the ALSFRS trajectories well, as can be seen from the fitted and residual plots in Figure 4. The scatterplot of fitted vs. observed ALSFRS scores showed that only a small proportion of the fitted ALSFRS scores deviated more than four points from the observed scores, and that the observed ALSFRS trajectories closely followed patient-specific linear functions. The scatterplot of the ALSFRS slope derived from the mixed-model (2) and the ALSFRS score ratio (Figure 4) showed that the ALSFRS slope $\hat{\beta}_1 + \hat{\gamma}_i$ is highly correlated with the ALSFRS score ratio but does not show many exactly-zero slopes, has a smaller number of positive slopes, and seems to be less variable. A direct comparison of the empirical cumulative distribution functions of the two slopes in Figure 3 also led to these conclusions.

Gordon et al. Gordon *et al.* (2010) reported that the mean ALSFRS trajectory over time is nonlinear. We also modelled deviations from a linear ALSFRS trajectory by allowing an additional fixed effect in quadratic time $\beta_2 t^2$ in model (2). This model depicted in the left part of Figure 5 (A) indeed suggested a nonlinear ALSFRS trajectory; however, we found a faster instead of a slower disease progression, as reported earlier (cf. Figure 1 in Gordon *et al.* 2010). The random slope parameters necessary for the definition of the ALSFRS slope between the linear and the quadratic model were highly correlated. The specific functional form of the mean ALSFRS trajectory therefore did not seem to be important for measuring the individual ALSFRS slopes, and we therefore obtained the ALSFRS slope for prediction purposes from the linear model (2).

[Figure 5 about here.]

Our results indicated that the ALSFRS score ratio suffers a high variability. Furthermore, the ALSFRS score ratio cannot be treated as a continuous variable, mainly because of some patients with a zero ALSFRS score ratio, which induces conceptual problems in many learning algorithms. With the ALSFRS slope, we defined an alternative estimator for the same theoretical slope parameter that describes the medical condition we are actually interested in and which is continuous and less variable, does not depend on the examination schedule of the patient, and can be fitted using a standard random intercept-random slope linear mixed-model for longitudinal data. Furthermore, the ALSFRS slope does not depend on the nine month prediction period but measures disease progression independently of time, taking all available information into account.

3.2. Prediction

The conditional random forest approach to ALSFRS slope prediction described here was ranked third out of 37 unique approaches submitted in the challenge. The mean RMSE and Pearson correlation of our conditional random forest predictions for the ALSFRS score ratio for the bootstrapped validation samples were 0.5208 and 0.4041, respectively. The first and second place winning teams performed better on average (RMSE of 0.5113 and 0.5152, DREAM and Prize4Life 2012). These numbers are conditional on the model being estimated on the fixed training and test samples. For the internal validation, we resampled from all patients available to the solvers and re-fitted the conditional inference forest 100 times, also capturing the variability of the fitted model (Hothorn *et al.* 2005). With the estimated variability from our simulation experiments, the differences in RMSE are most probably due to random error and not to significantly different prediction accuracies in an unconditional way (Figure 6).

[Figure 6 about here.]

From a conditional random forest fitted to all patients in the training and test data sets, we computed variable importances to guide model interpretation; the results are displayed in Figure 7. The most important variable was the ALSFRS slope up to 92 days, i.e., the lagged target variable. This is not surprising for longitudinal data under a linearity assumption. In addition, all ten ALSFRS-item-specific slopes received a high variable importance, which indicated that the ALSFRS sum score might not contain all that is to be known about the disease progression. Also the region of disease onset seemed to play an important role in the model. It should be noted that the onset variable itself was important, but this was an artefact of the data because the data contain almost no information about patients with very slow progression that were included very early (i.e., with small onset values); therefore, large absolute onset corresponded to slow progression. The remaining variables seemed to be only of marginal or even zero importance for the accuracy of this prediction model.

The association between the most important variables and the predicted ALSFRS slope is depicted by means of partial dependency plots in Figure 8. Positive ALSFRS slopes in the first three months, i.e., no measureable disease progression, was associated with smaller predicted slopes. A large variability of the ALSFRS scores in the first three months led to larger predicted slopes. Bulbar onset was associated with steeper slopes compared to limb onset; this confirms findings reported on earlier (Qureshi *et al.* 2006; Gordon *et al.* 2010).

[Figure 7 about here.]

[Figure 8 about here.]

4. Discussion

Given the high expectations the challenge organisers might have had, the conclusion we can draw from our approach to ALSFRS slope prediction is somewhat limited: at least with the data on which this challenge was based, the best predictor for future ALSFRS slope is the

past ALSFRS slope, a fact already well known (e.g. Kimura *et al.* 2006). Except for onset site, we could not find any baseline characteristics or any other variable that promises to be a strong candidate for predicting ALS disease progression. However, two interesting findings are worth further investigation. First, in addition to the initial ALSFRS slope, the range of the ALSFRS scores read in the first three months after study entry was also a strong predictor variable, with larger ALSFRS score variability being linked to a faster disease progression. In itself, the variability might not be important but our interpretation is that the increased variability might be associated with an unknown factor that has quite a substantial impact on disease progression. Second, the item-specific slopes of all ten items that define the ALSFRS score had a high variable importance and therefore help to improve the prediction quality. Especially the item "speech" seemed to offer additional information about ALSFRS disease progression. This finding is in line with a recent report on the multidimensionality of the ALSFRS-R score (Franchignoni *et al.* 2013).

As a consequence, a possible improvement of our model would be the application of a polytomous Rasch model for longitudinal observations, in which a latent parameter describing disease progression can be directly modelled for all ALSFRS items. This model would also overcome the rather unrealistic normal assumption on the residuals on which our analysis is based. One conceptual problem ignored in our analysis was that we cannot assume the independence of loss of follow-up and unobserved ALSFRS scores. Thus, the missing at random (MAR) assumption is not justified, and the mixed-model results may be biased.

Beside these more technical shortcomings of our approach, the generalisability of our findings and also the clinical relevance might be limited due to experimental design and data collection. The models discussed here try to predict ALSFRS score trajectories, which only partially describe disease progression. Furthermore, due to confidentiality issues, the names of the studies the patients participated in were not published and thus we have to assume that the study population is quite difference from the general ALS population (Chió *et al.* 2011). This might explain why known prognostic factors, such as age and forced vital capacity, do not play an important role in our models.

After the challenge was closed on 20 October 2012, the PRO-ACT database was made available for interested researchers, now also including records of more than 6000 additional patients. According to the sample sizes of the studies reported on in (Chiò *et al.* 2009), this is the largest collection of ALS patient data available for the identification of prognostic factors and the development of prediction models. The analysis of this database with methods similar to the approach discussed here thus promises to lead to further insights into how we can come up with better prognostic models for ALS disease progression.

Acknowledgements

Neta Zach of Prize4Life constantly answered our questions on the PRO-ACT database during the challenge and later during manuscript preparation and we would like to thank her for her invaluable support. Karen A. Brune helped us to improve the language. We thank two anonymous referees for their helpful comments on the initial manuscript.

Disclosure of Interests

TH was awarded the third prize of $10000 for his contribution to the challenge. Both authors have no further financial interests in the results reported here.

Contribution to Authorship

TH analysed the data and drafted the manuscript, both authors interpreted the results, revised the description of methods and results, and approved this version of the manuscript.

References

Bates D, Maechler M, Bolker B (2013). ***lme4****: Linear Mixed-effects Models Using S4 Classes*. R package version 0.999999-2, URL http://CRAN.R-project.org/package=lme4.

Breiman L (2001). "Random Forests." *Machine Learning*, **45**(1), 5–32.

Brettschneider J, Toledo JB, Deerlin VMV, Elman L, McCluskey L, Lee VMY, Trojanowski JQ (2012). "Microglial Activation Correlates With Disease Progression and Upper Motor Neuron Clinical Symptoms in Amyotrophic Lateral Sclerosis." *PLoS ONE*, **7**(6), e39216. doi:10.1371/journal.pone.0039216.

Brooks BR, Lewis D, Rawling J, Sanjak M, Belden D, Hakim H, De Tan Y, Sufit R, Gaffney J, Depaul R (1994). "The Natural History of Amyotrophic Lateral Sclerosis." In AC Williams (ed.), "Motor Neuron Disease," volume XVI, pp. 131–170. Chapman & Hall Medical, London, UK.

Brooks BR, Sanjak M, Ringel S, England J, Brinkmann J, Pestronk A, Florence J, Mitsumoto H, Szirony K, Wittes J (1996). "The Amyotrophic Lateral Sclerosis Functional Rating Scale: Assessment of Activities of Daily Living in Patients With Amyotrophic Lateral Sclerosis." *Archives of Neurology*, **53**(2), 141–147.

Cedarbaum JM, Stambler N, Malta E, Fuller C, Hilt D, Thurmond B, Nakanishi A (1999). "The ALSFRS-R: a Revised ALS Functional Rating Scale That Incorporates Assessments of Respiratory Function." *Journal of the Neurological Sciences*, **169**(1), 13–21.

Chió A, Canosa A, Gallo S, Cammarosano S, Moglia C, Fuda G, Calvo A, M Gabriele for the PARALS Group (2011). "ALS Clinical Trials: Do Enrolled Patients Accurately Represent the ALS Population?" *Neurology*, **77**(15), 1432–1437. doi:10.1212/WNL.0b013e318232ab9b.

Chiò A, Logroscino G, Hardiman O, Swingler R, Mitchell D, Beghi E, Traynor BG on behalf of the Eurals Consortium (2009). "Prognostic Factors in ALS: A Critical Review." *Amyotrophic Lateral Sclerosis*, **10**(5-6), 310–323. doi:10.3109/17482960802566824.

Diggle PJ, Heagerty PJ, Liang K, Zeger SL (2002). *Analysis of Longitudinal Data*. Oxford University Press, Oxford, UK, 2nd edition.

DREAM, Prize4Life (2012). "The DREAM Phil Bowen ALS Prediction Prize4Life Challenge." Accessed 2013-05-24, URL http://www.the-dream-project.org/result/dream-phil-bowen-als-prediction-prize4life-challenge.

DREAM Project (2013). "DREAM Project Webpage." Accessed 2013-05-24, URL http://www.the-dream-project.org/.

Franchignoni F, Mora G, Giordano A, Volanti P, Chió A (2013). "Evidence of Multidimensionality in the ALSFRS-R Scale: a Critical Appraisal on Its Measurement Properties Using Rasch Analysis." *Journal of Neurology, Neurosurgery & Psychiatry*. doi:10.1136/jnnp-2012-304701. Online first.

Fujimura-Kiyono C, Kimura F, Ishida S, Nakajima H, Hosokawa T, Sugino M, Hanafusa T (2011). "Onset and Spreading Patterns of Lower Motor Neuron Involvements Predict Survival in Sporadic Amyotrophic Lateral Sclerosis." *Journal of Neurology, Neurosurgery & Psychiatry*, **82**(11), 1244–1249. doi:10.1136/jnnp-2011-300141.

Gordon PH, Cheng B, Salachas F, Pradat PF, Bruneteau G, Corcia P, Lacomblez L, Meininger V (2010). "Progression in ALS Is Not Linear But Is Curvilinear." *Journal of Neurology*, **257**(10), 1713–1717. doi:10.1007/s00415-010-5609-1.

Hapfelmeier A, Hothorn T, Ulm K, Strobl C (2012). "A New Variable Importance Measure for Random Forests with Missing Data." *Statistics and Computing*. doi:10.1007/s11222-012-9349-1.

Hothorn T, Hornik K, Strobl C, Zeileis A (2013). ***party**: A Laboratory for Recursive Part(y)itioning*. R package version 1.0-6, URL http://CRAN.R-project.org/package=party.

Hothorn T, Hornik K, Zeileis A (2006). "Unbiased Recursive Partitioning: A Conditional Inference Framework." *Journal of Computational and Graphical Statistics*, **15**(3), 651–674. doi:10.1198/106186006X133933.

Hothorn T, Leisch F, Zeileis A, Hornik K (2005). "The Design and Analysis of Benchmark Experiments." *Journal of Computational and Graphical Statistics*, **14**(3), 675–699. doi:10.1198/106186005X59630.

InnoCentive (2013). "InnoCentive Webpage." Accessed 2013-05-24, URL http://www.innocentive.com/.

Kimura F, Fujimura C, Ishida S, Nakajima H, Furutama D, Uehara H, Shinoda K, Sugino M, Hanafusa T (2006). "Progression Rate of ALSFRS-R At Time of Diagnosis Predicts Survival Time in ALS." *Neurology*, **66**(2), 265–267. doi:10.1212/01.wnl.0000194316.91908.8a.

Kollewe K, Mauss U, Krampfl K, Petri S, Dengler R, Mohammadi B (2008). "ALSFRS-R Score and Its Ratio: A Useful Predictor for ALS-progression." *Journal of the Neurological Sciences*, **275**(1-2), 69–73. doi:10.1016/j.jns.2008.07.016.

Küffner R, Zach N, Norel R, Hawe J, Schoenfeld D, Wang L, Li G, Fang L, Mackey L, Hardiman O, Cudkowicz M, Sherman A, DREAM 7 ALS Prediction Prize Consortium, Stolovitzky G, Leitner M (2013). "Predicting ALS Progression: Insights Into the Disease From a Crowdsourcing Effort." Unpublished manuscript.

Paganoni S, Zhang M, Zárate AQ, Jaffa M, Yu H, Cudkowicz ME, Wills AM (2012). "Uric Acid Levels Predict Survival in Men With Amyotrophic Lateral Sclerosis." *Journal of Neurology*, **259**(9), 1923–1928. doi:10.1007/s00415-012-6440-7.

Prize4Life (2013). "Prize4Life Webpage." Accessed 2013-05-24, URL http://www.prize4life.org/.

Prize4Life, Neurological Clinical Research Institute, Massachusetts General Hospital (2013a). "A Guide to the PRO-ACT Database." Accessed 2013-05-24, URL https://nctu.partners.org/ProACT/Home/HowToUse.

Prize4Life, Neurological Clinical Research Institute, Massachusetts General Hospital (2013b). "Pooled Resource Open-Access ALS Clinical Trials Database." Accessed 2013-05-24, URL https://nctu.partners.org/ProACT/.

Prize4Life, Neurological Clinical Research Institute, Massachusetts General Hospital (2013c). "Pooled Resource Open-Access ALS Clinical Trials Database–ALS Prize Data." Accessed 2013-11-04, URL https://nctu.partners.org/ProACT/data.

Qureshi MM, Hayden D, Urbinelli L, Ferrante K, Newhall K, Myers D, Hilgenberg S, Smart R, Brown RH, Cudkowicz ME (2006). "Analysis of Factors That Modify Susceptibility and Rate of Progression in Amyotrophic Lateral Sclerosis (ALS)." *Amyotrophic Lateral Sclerosis*, **7**(3), 173–182. doi:10.1080/14660820600640596.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Reich-Slotky R, Andrews J, Cheng B, Buchsbaum R, Levy D, Kaufmann P, Thompson JLP (2013). "Body Mass Index (BMI) as Predictor of ALSFRS-R Score Decline in ALS Patients." *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, **14**(3), 212–216. doi:10.3109/21678421.2013.770028.

Rowell JG, Walters DE (1976). "Analysing Data With Repeated Observations on Each Experimental Unit." *Journal of Agricultural Science*, **87**, 423–432.

Strobl C, Boulesteix AL, Zeileis A, Hothorn T (2007). "Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution." *BMC Bioinformatics*, **8**, 25. doi:10.1186/1471-2105-8-25.

Strobl C, Malley J, Tutz G (2009). "An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests." *Psychological Methods*, **14**(4), 323–348. doi:10.1037/a0016973.

Wishart J (1938). "Growth-rate Determinition in Nutrition Studies With the Bacon Pig, and their Analysis." *Biometrika*, **30**, 16–28.

Zoccolella S, Beghi E, Palagano G, Fraddosio A, Guerra V, Samarelli V, Lepore V, Simone IL, Lamberti P, Serlenga L, Logroscino G for the SLAP Registry (2008). "Analysis of Survival and Prognostic Factors in Amyotrophic Lateral Sclerosis: a Population Based Study." *Journal of Neurology, Neurosurgery & Psychiatry*, **79**(1), 33–37. doi:10.1136/jnnp.2007.118018.

Affiliation:

Torsten Hothorn
Institut für Sozial- und Präventivmedizin, Abteilung Biostatistik
Universität Zürich
Hirschengraben 84
CH-8001 Zürich, Switzerland
E-mail: Torsten.Hothorn@uzh.ch
URL: http://www.biostat.uzh.ch/aboutus/people/hothorn.html

Hans H. Jung
Klinik für Neurologie
UniversitätsSpital Zürich
Frauenklinikstrasse 26
CH-8091 Zürich, Switzerland

List of Figures

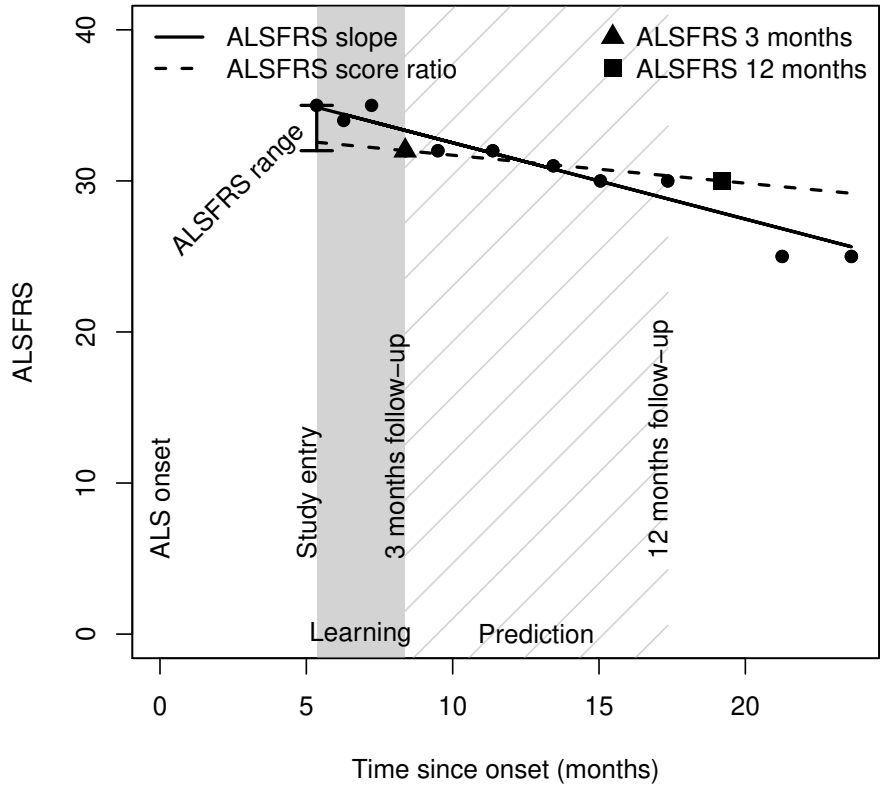

Figure 1: ALSFRS trajectory and schedule for patient number 29253. All information gathered in the three month period after study entry (learning period; shaded area) was available for computing predictions for the ALSFRS trajectory between three and twelve months (prediction period; grey hatched area). The filled triangle and square symbols indicate the two ALSFRS readings that define the ALSFRS score ratio (dashed line). The ALSFRS slope (solid line) was obtained from the linear mixed-model (2). The ALSFRS range was computed for the ALSFRS examinations in the first three months.

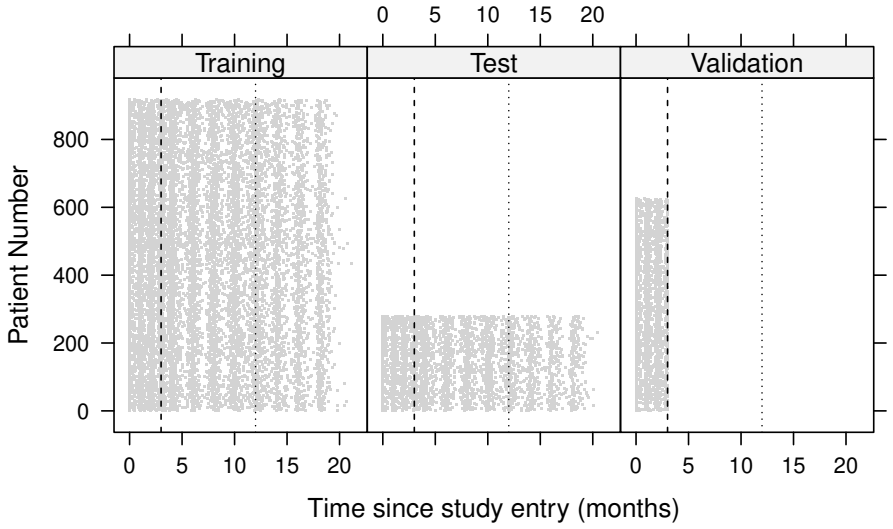

Figure 2: Patient data overview. For the three samples the challenge was based on (training, test, and validation), the examination dates for each patient (in months since study entry) are visualised by dots in one row. Data obtained during the first three months (learning period, dashed vertical line) for patients in the training and test samples was used as input for fitting the models whereas the disease progression up to 12 month (prediction period, dotted line) defined the outcome. The aim of the challenge was to predict the undisclosed ALSFRS score trajectories between three and 12 months for the patients in the validation sample.

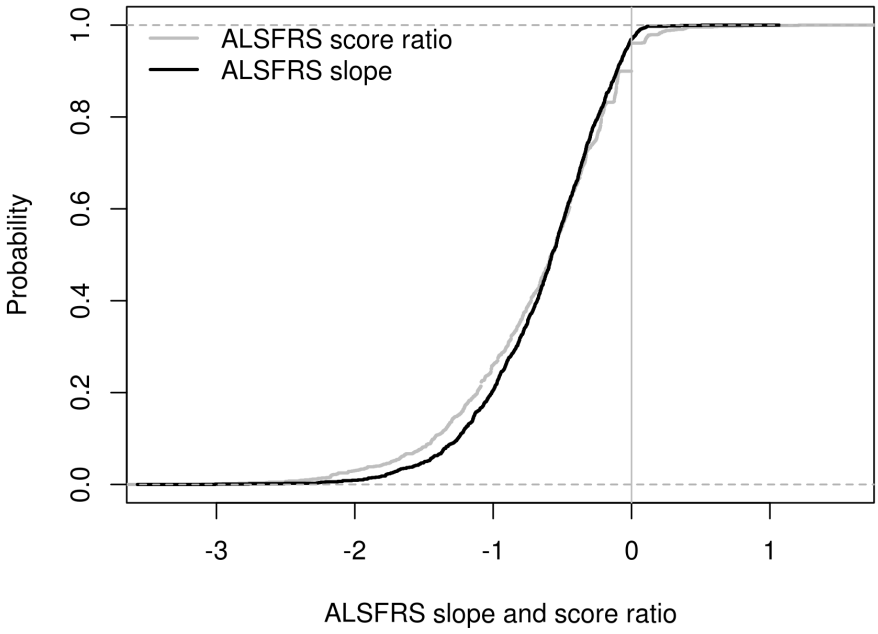

Figure 3: Empirical distribution function of the ALSFRS slope (black line) and ALSFRS score ratio (grey line) for the 1197 patients in the training and test samples.

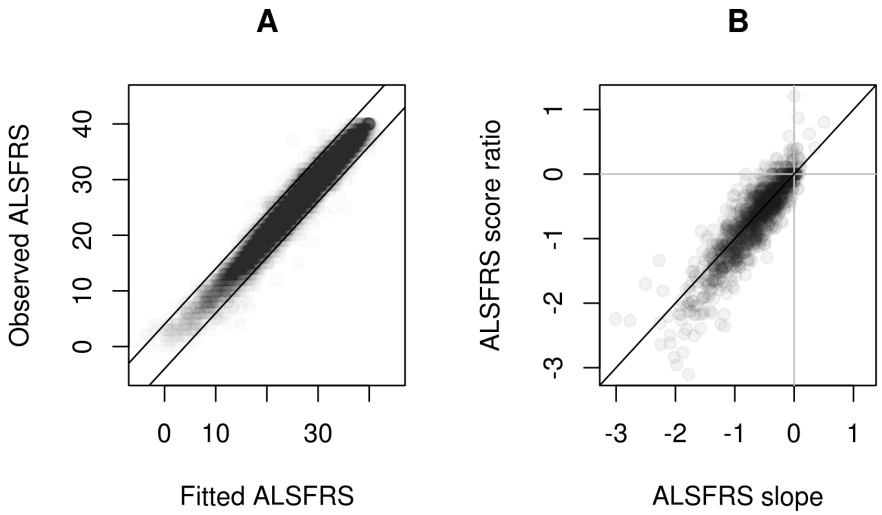

Figure 4: A) Comparison of fitted and observed ALSFRS scores based on the linear mixed-model (2). Lines indicate ±4 ALSFRS scores. B) Scatterplot of the ALSFRS score ratio and the ALSFRS slope obtained from the mixed-model.

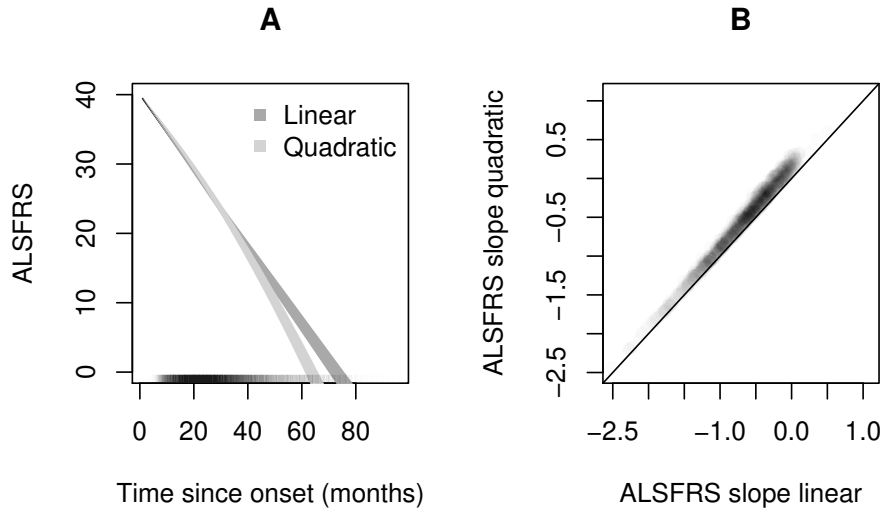

Figure 5: Comparison of A) linear and quadratic models for the mean ALSFRS trajectory over time. Confidence bands for the expected ALSFRS over time obtained from the two models are displayed. Rugs indicate times of patient examinations. B) Random slopes obtained from the linear and quadratic model.

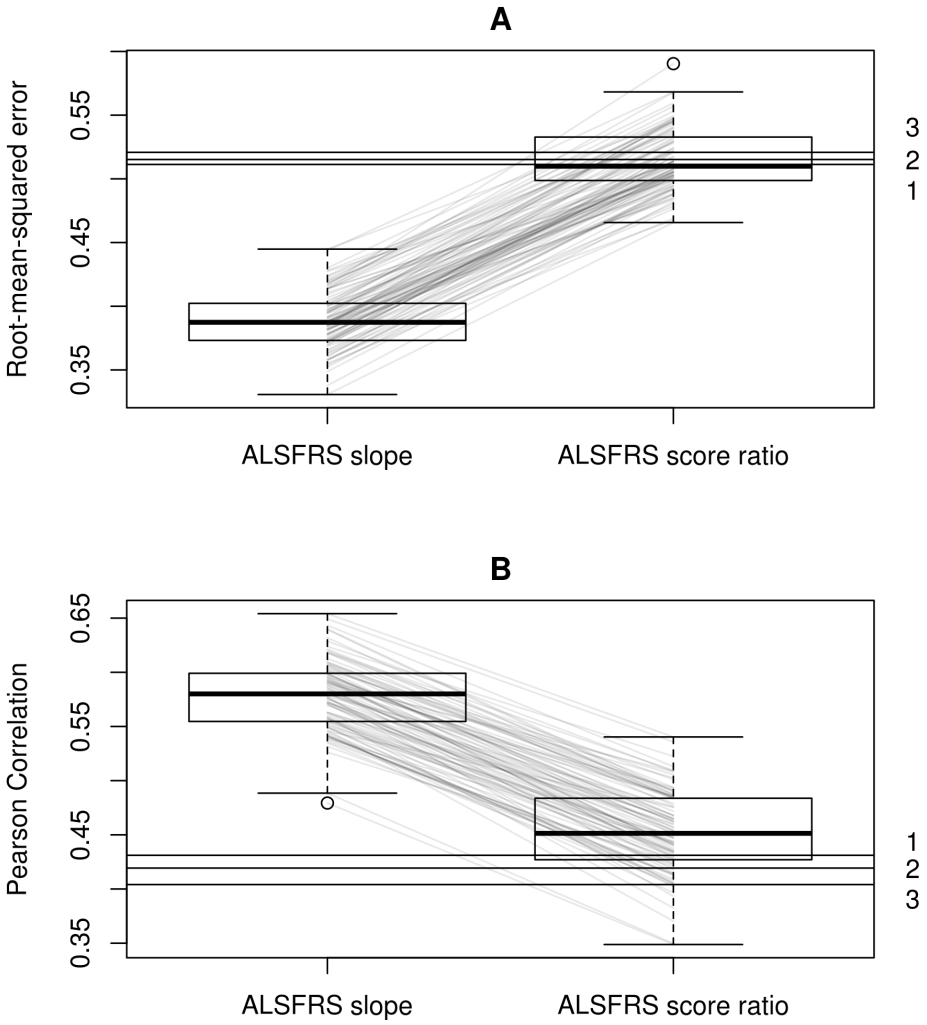

Figure 6: A) Root-mean-squared error and B) correlation of conditional random forest fitted to the ALSFRS slope for predicting the ALSFRS slope (left boxplots), and conditional random forest fitted to the ALSFRS slope for predicting the ALSFRS score ratio (right boxplots). The black horizontal lines indicate the validation sample prediction performance of the three winning teams: 3 corresponds to our method described here.

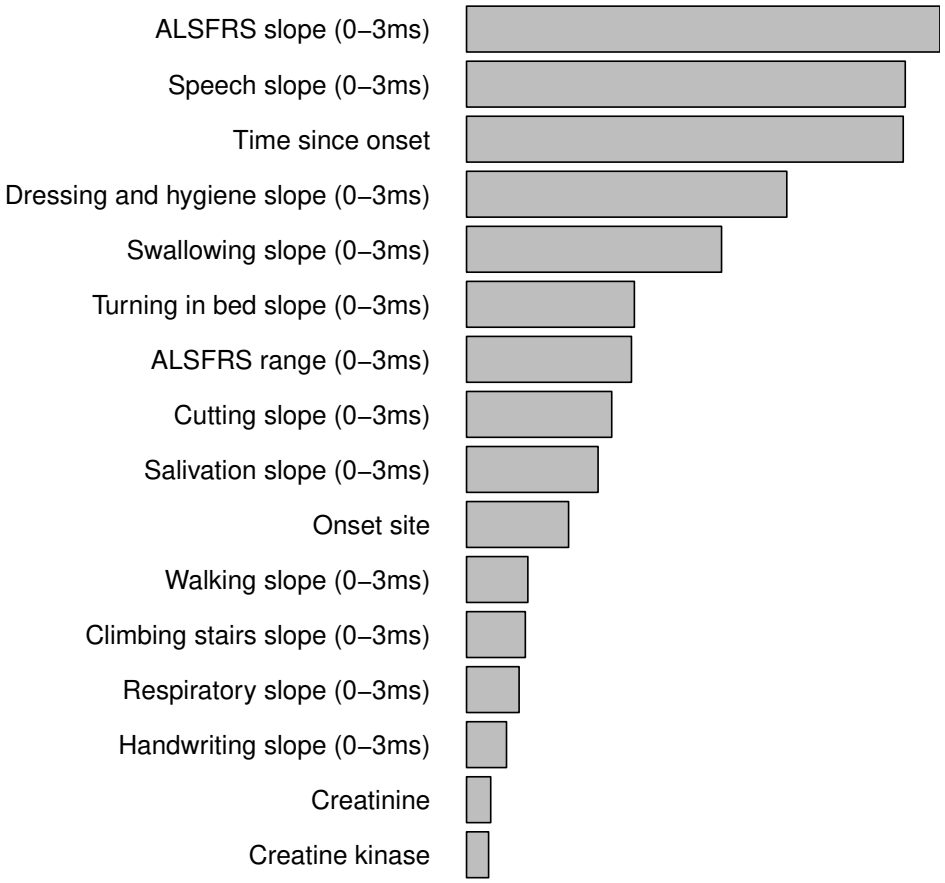

Figure 7: Permutation variable importances of conditional random forest fitted to the ALS-FRS slopes on the training and test samples. The longer the bar, the more important the corresponding variable. The remaining variables not shown here, especially the large set of laboratory parameters, received a very low permutation importance that indicates a very small or even nonexistent impact on disease progression. (0-3 ms) refers to the initial month observation period.

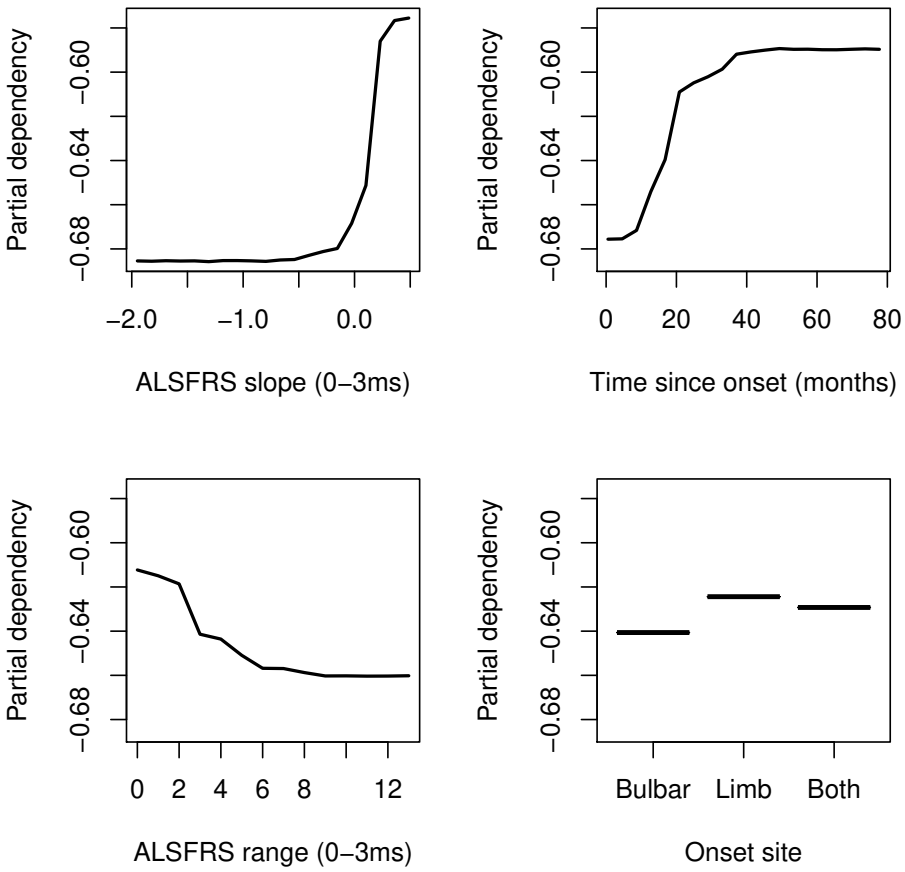

Figure 8: Partial dependency plots for selected important variables. On the ordinate, partial dependencies as mean predicted ALSFRS slopes are given.

List of Tables

		Training	Test	Validation
Sex				
	Female	53	182	107
	Male	91	303	222
	na	135	433	296
Race				
	Caucasian	0	8	5
	Asian	4	14	9
	Black/African Am.	272	886	598
	na	3	10	13
Onset site				
	Bulbar	59	203	121
	Limb	218	711	500
	Limb and bulbar	2	4	4
Age (years)		53 (44–62)	55 (47–64)	56 (47–64)
Height (cm)		170 (164–178)	170 (163–178)	172 (165–179)
Weight (kg)		74 (63–86)	73 (60–85)	71 (60–84)
Number of visits		12 (11–12)	12 (11–12)	4 (3–4)
3-12 month period (days)		277 (258–305)	279 (260–303)	na

Table 1: Patient characteristics of the three patient samples. For numeric variables, the median along with the 1st and 3rd quartile (in parentheses) are given. Number of visits refers to the number of patient examinations, and 3-12-month period is the actual length of the prediction period defining the ALSFRS score ratio. For the validation data, examinations up to 92 days after study entry were made available to participating solvers after the challenge was closed; data from three months onwards was not published.